Multiparent populations & R/qtl2

Karl Broman

Biostatistics and Medical Informatics
University of Wisconsin – Madison
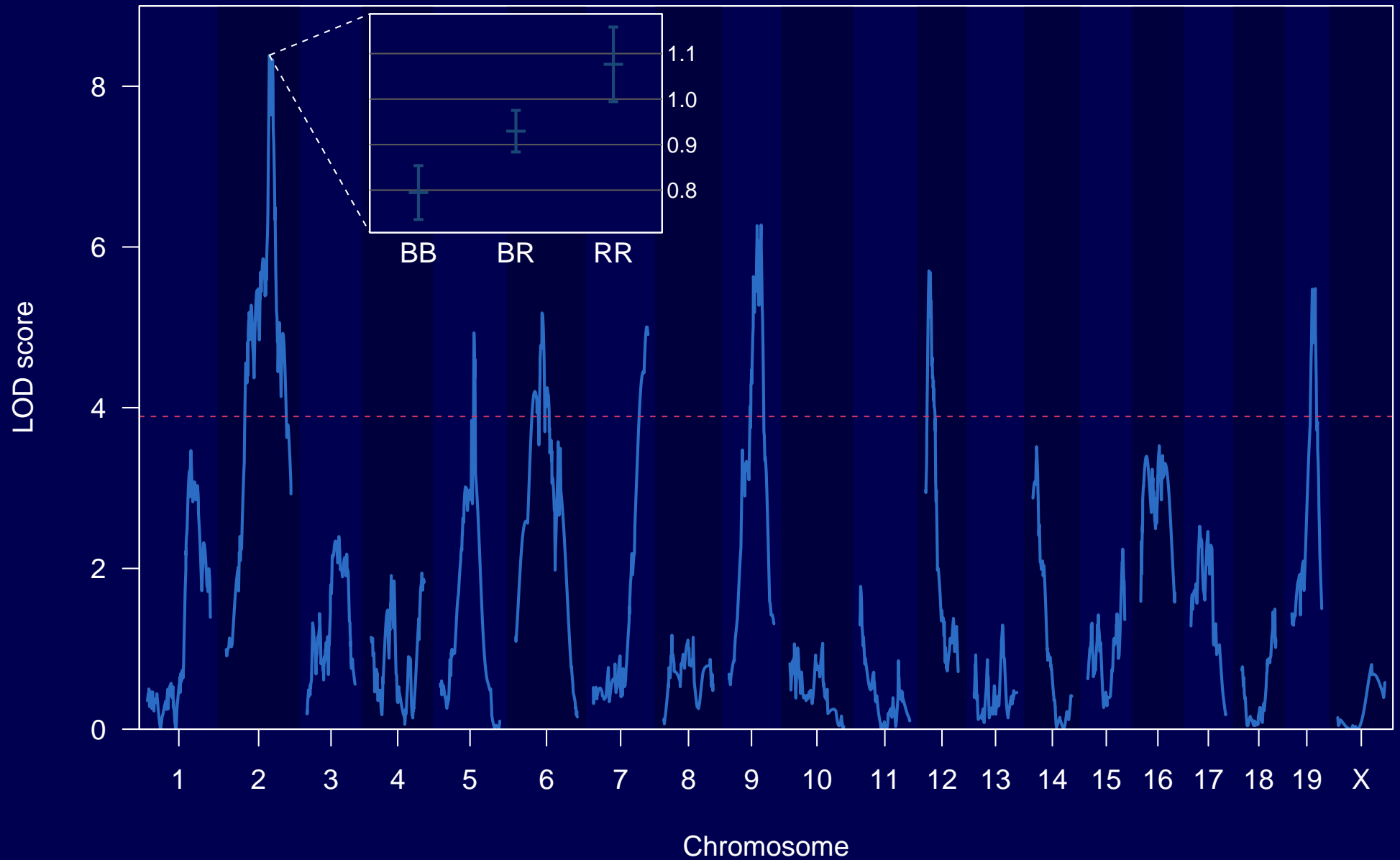
kbroman.org/qtl2

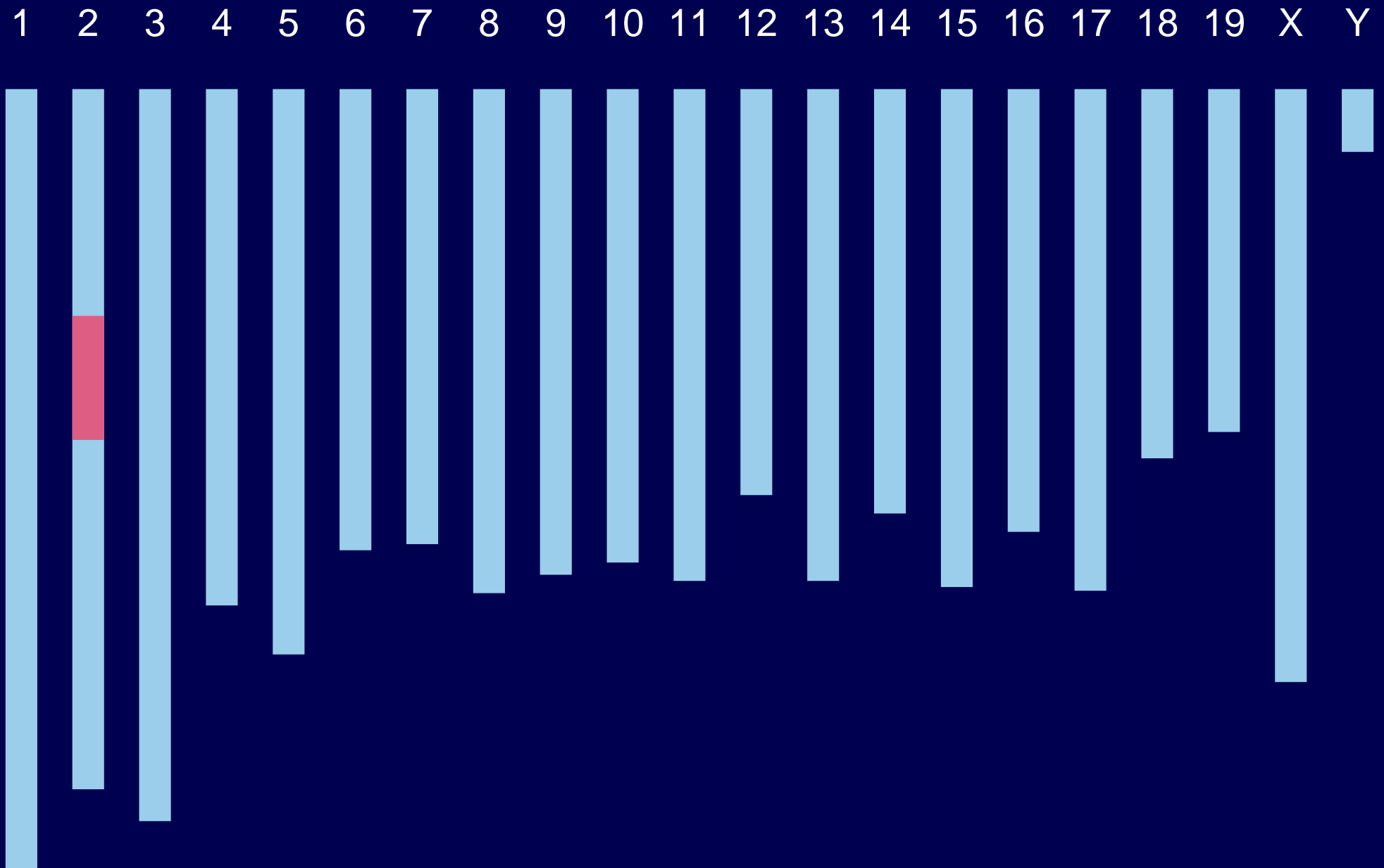kbroman.org

github.com/kbroman

@kwbroman

QTL mapping

Congenic line / NIL

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y
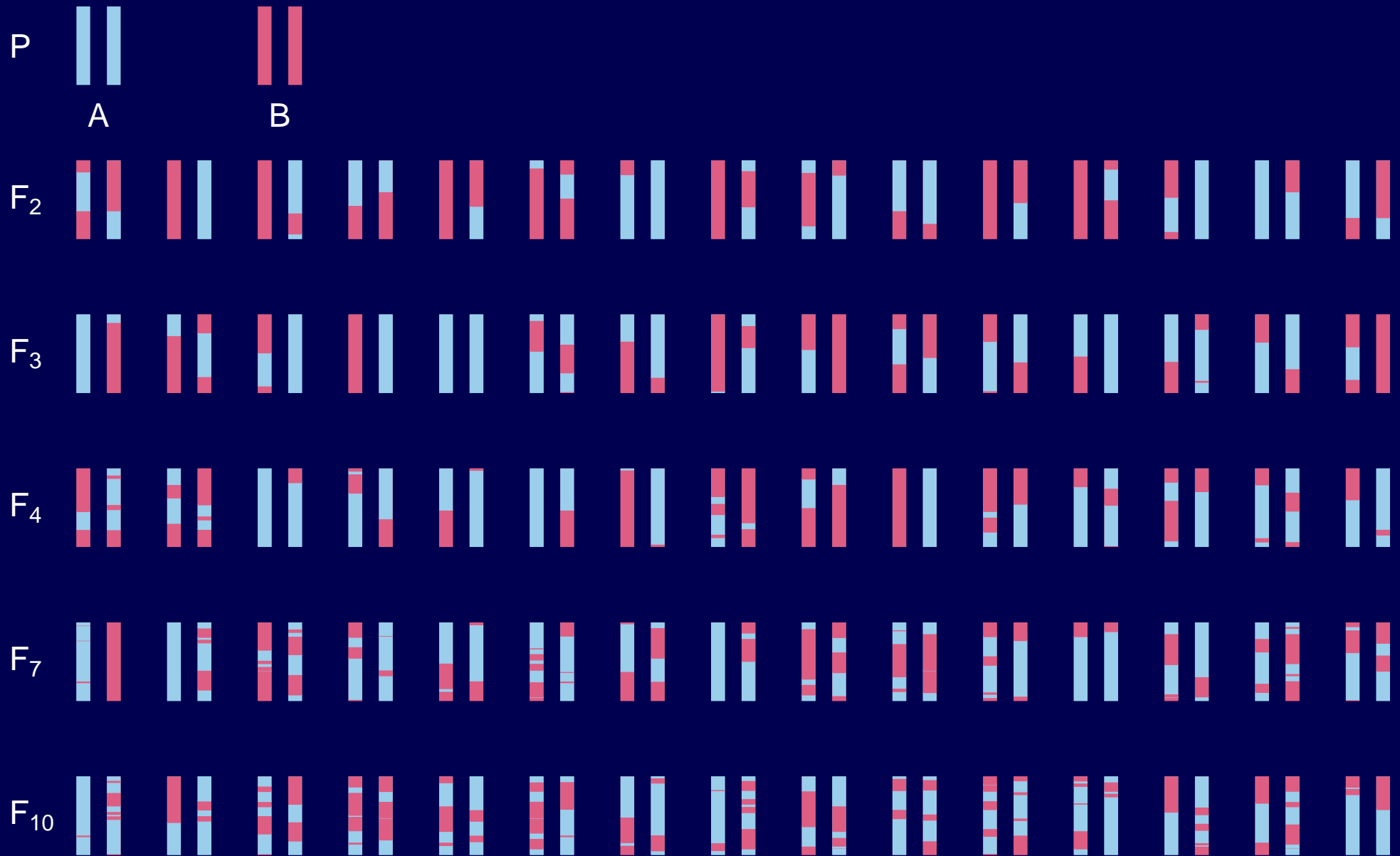
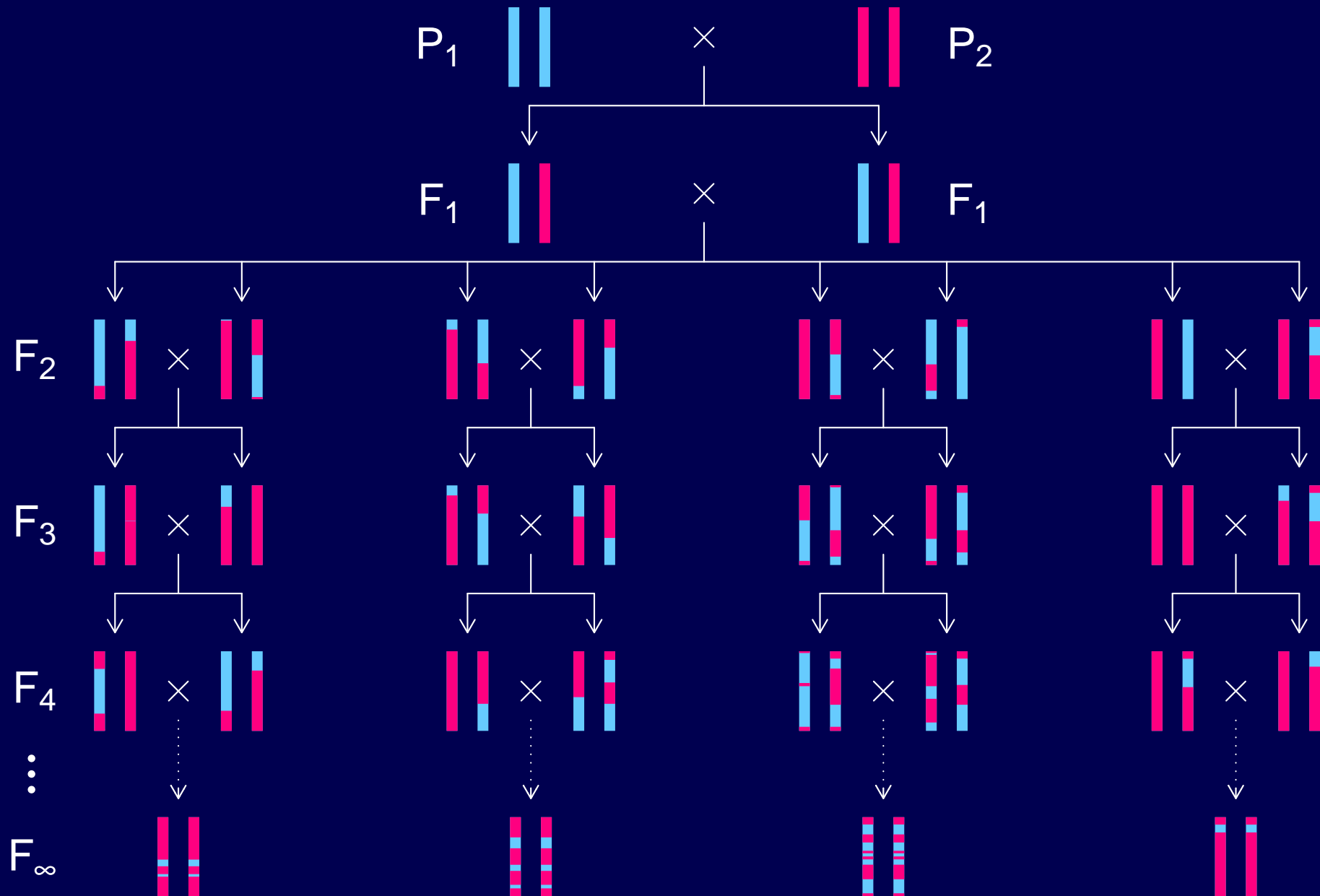# Improving precision

- more recombinations

- more individuals

- more precise phenotypes

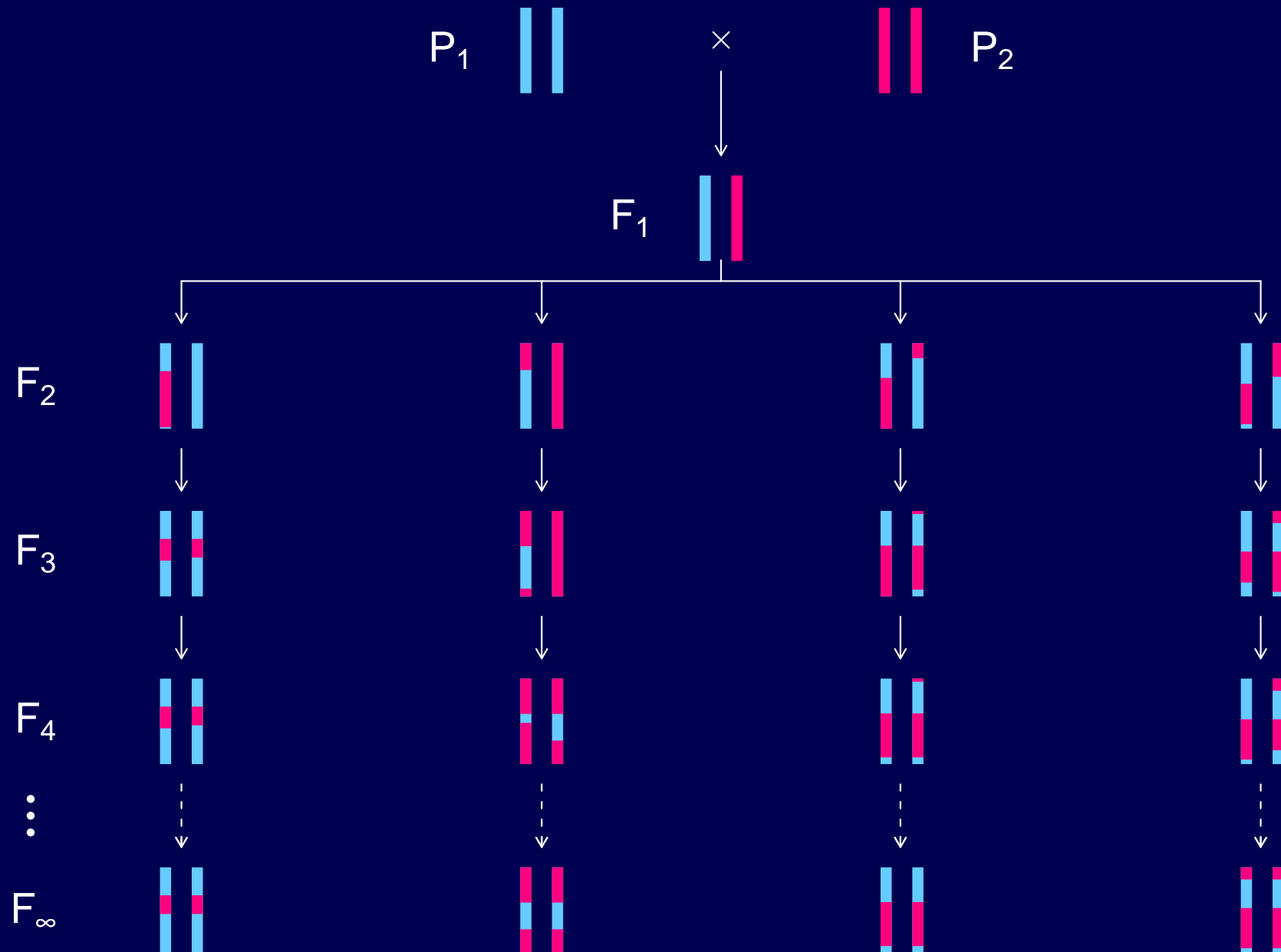- lower-level phenotypes

  transcripts, proteins, metabolites

Recombinant inbred lines
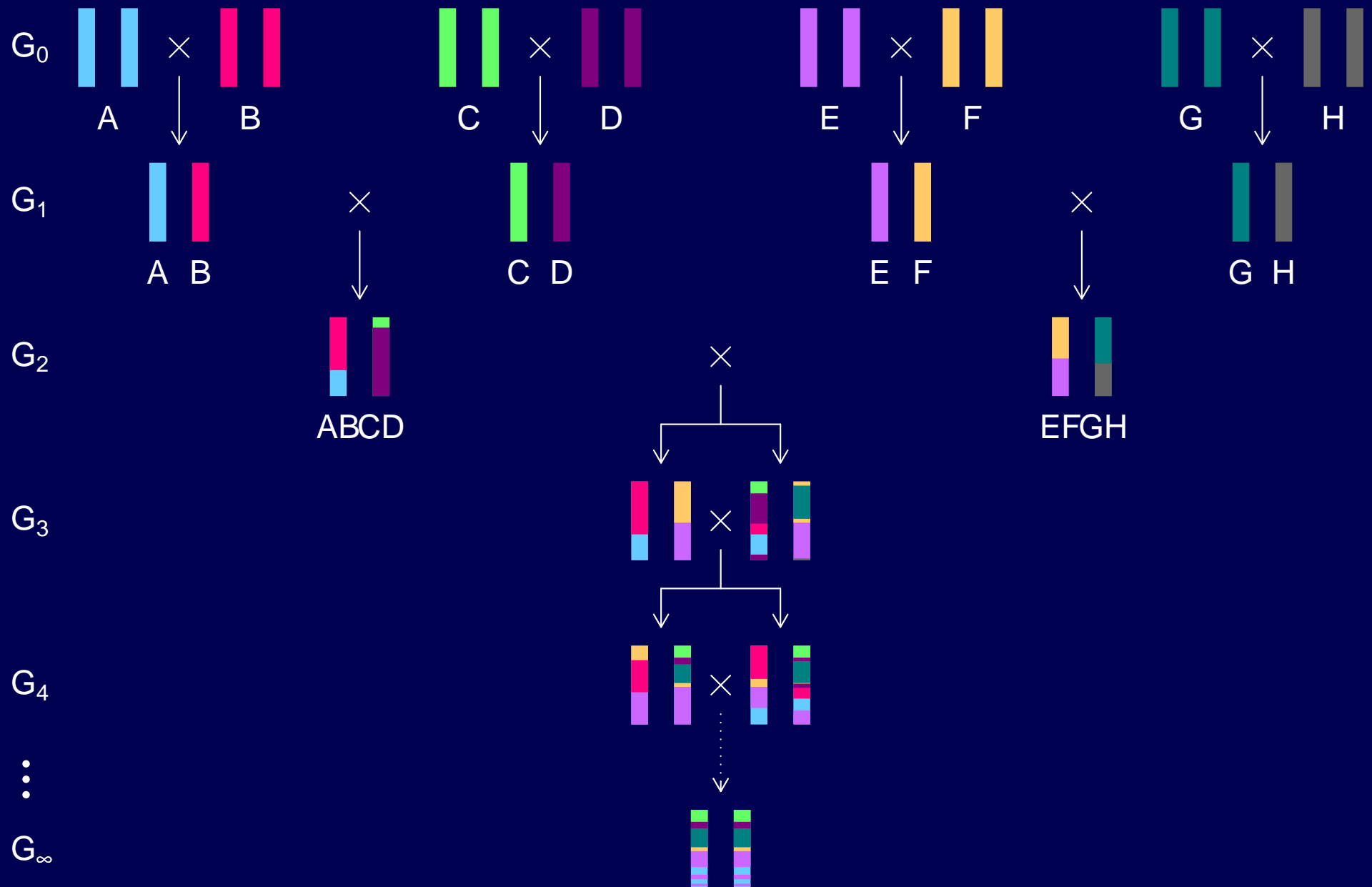
Recombinant inbred lines
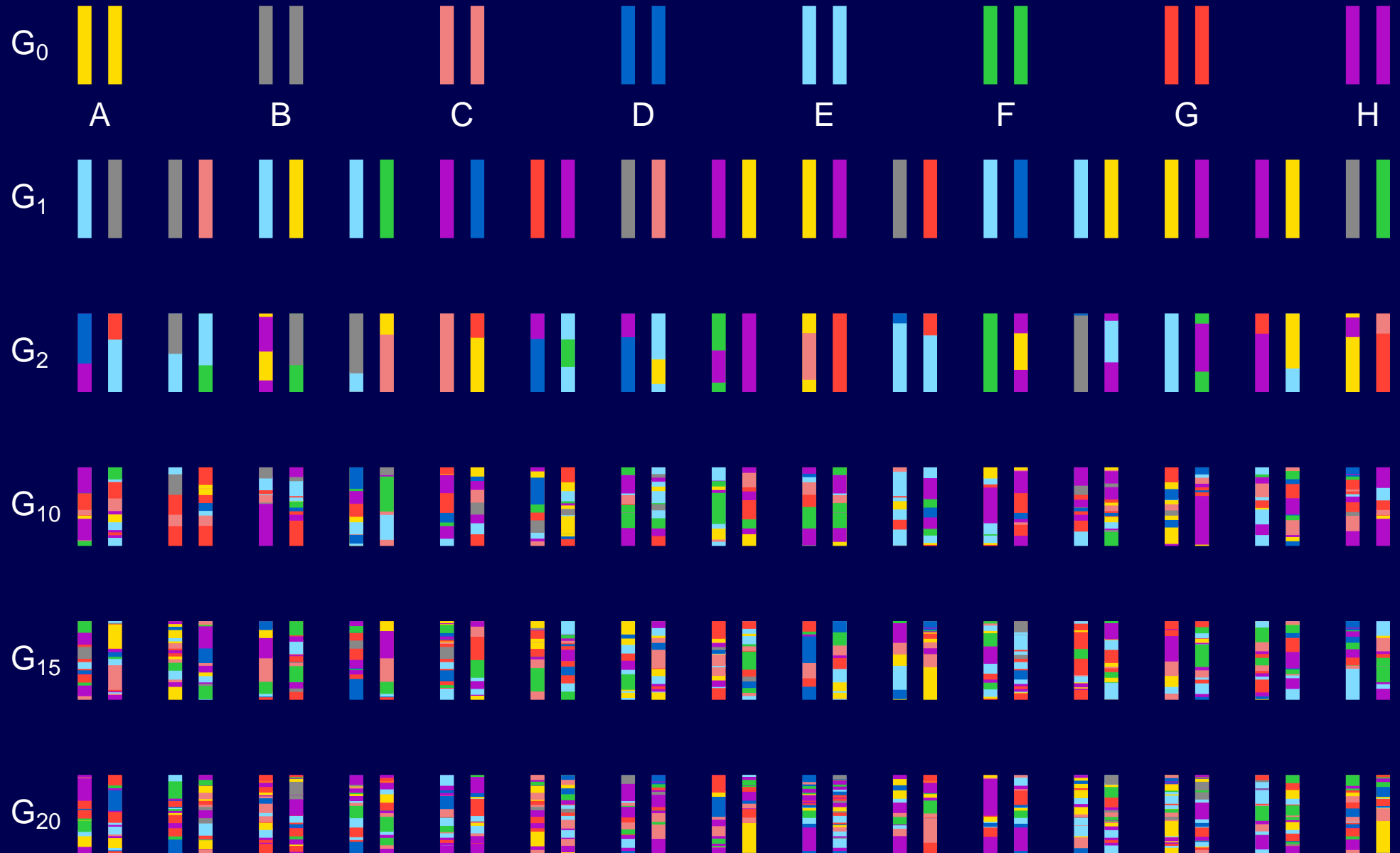
Collaborative Cross/MAGIC

G_0 A × B C × D E × F G × H

G_1 A B × C D E F × G H

G_2 ABCD × EFGH

G_3

G_4

G_∞

Heterogeneous stock

Why R/qtl2?

- High-dimensional data

  genotypes and phenotypes

- More diverse crosses

  especially multi-parent populations

- Linear mixed models

  especially in DO/HS/AIL

R/qtl $\rightarrow$ R/qtl2

- See `kbroman.org/qtl2/assets/vignettes/rqtl_diff.html`

- New data file formats

- New data structures

- New function names

  `read.cross()` $\rightarrow$ `read_cross2()`

  `calc.genoprob()` $\rightarrow$ `calc_genoprob()`

  `scanone()` $\rightarrow$ `scan1()`

- Different treatment of intermediate calculations

- Use of individual IDs for aligning data

- Order of args when subsetting cross objects

  `cross[chr,ind]` $\rightarrow$ `cross2[ind,chr]`

$\rightarrow$ R

- `convert2cross2()`

- `summary(), n_ind(), n_mar(),...`

- `insert_pseudomarkers()`

- `calc_genoprob()`
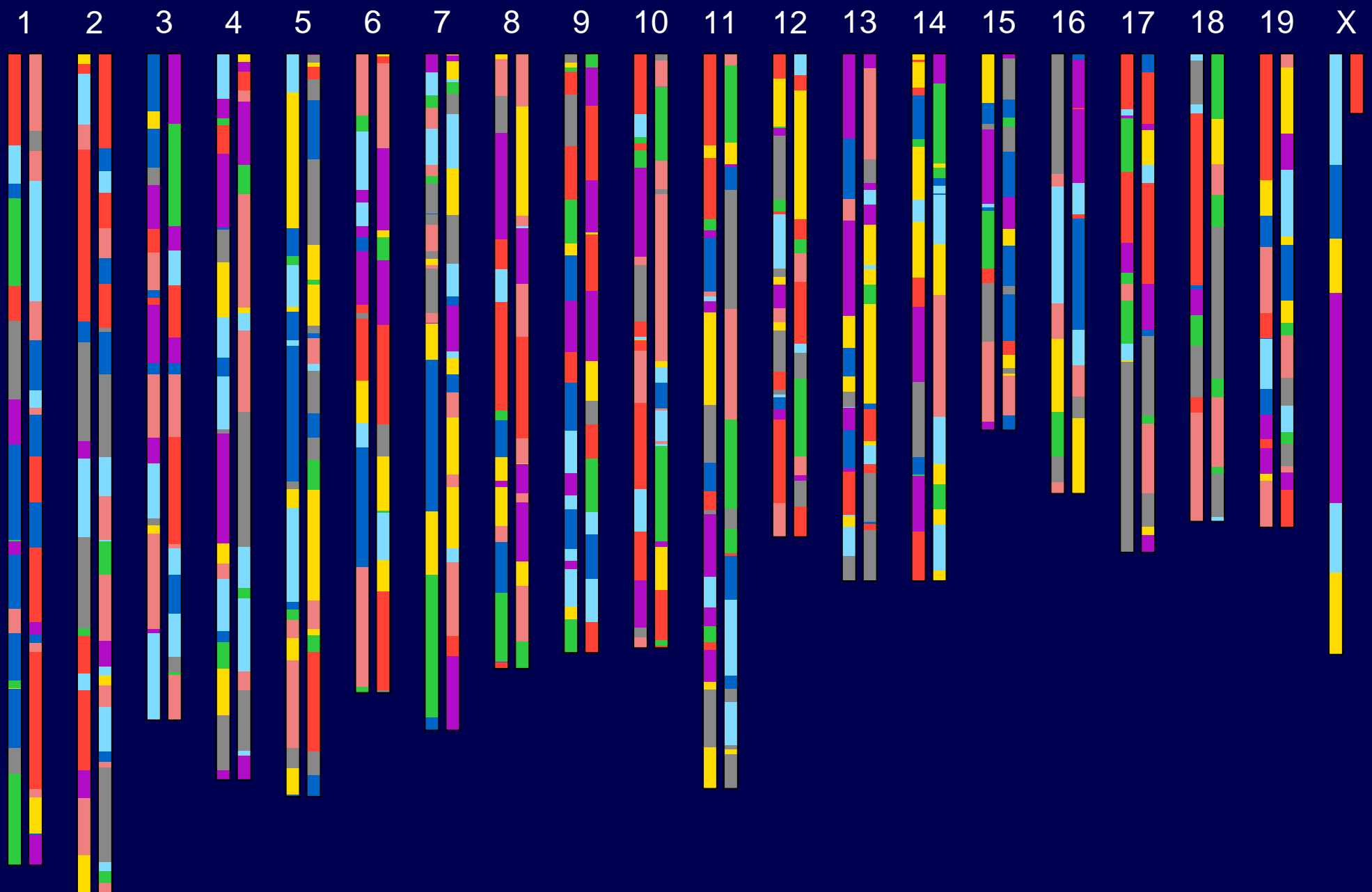
- `scan1()`

- `find_peaks()`

Linear mixed models

$$y_i = \mu + \sum_k \beta_k q_{ik} + \epsilon_i \qquad \epsilon_i \sim \mathsf{N}(0, \sigma_e^2)$$

$$= \mu + \eta_i + \epsilon_i \qquad \eta_i \sim \mathsf{N}(0, \sigma_p^2)$$
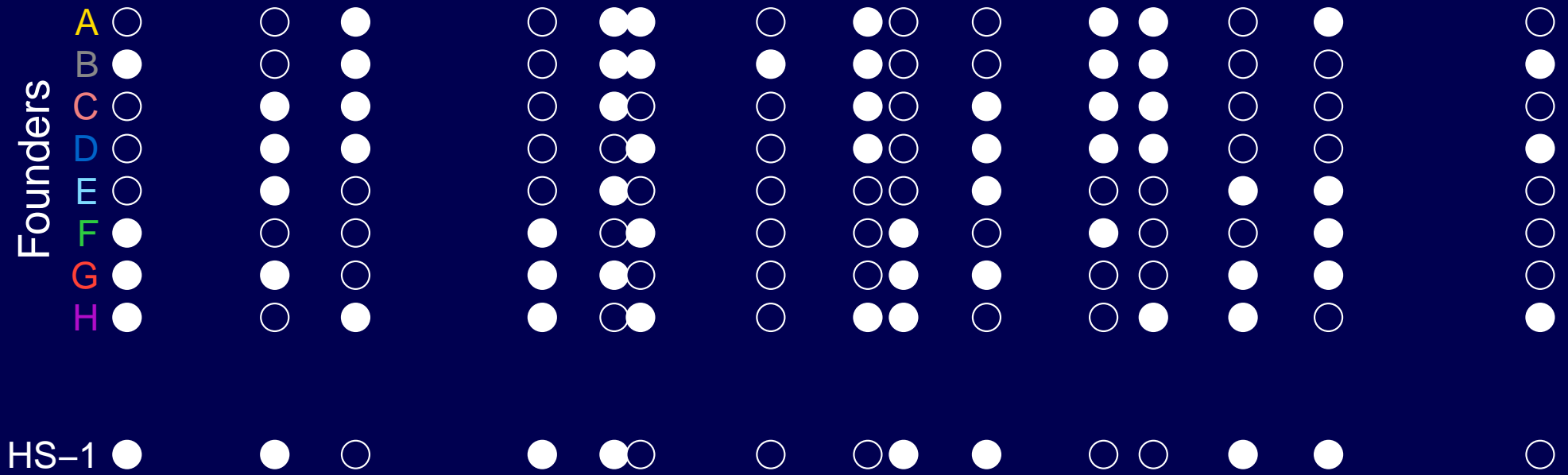
$$\mathsf{cov}(\eta_i, \eta_j) = \sigma_p^2 \, (2k_{ij})$$

→ R

- `calc_kinship()`

- `scan1()`

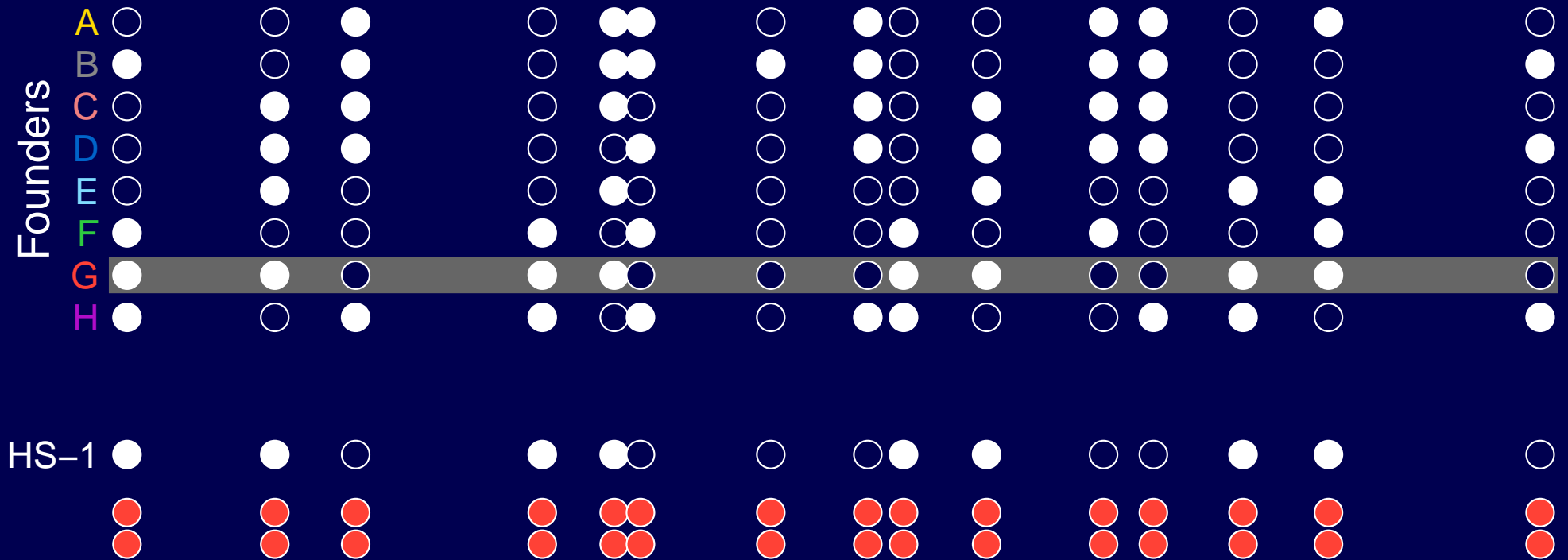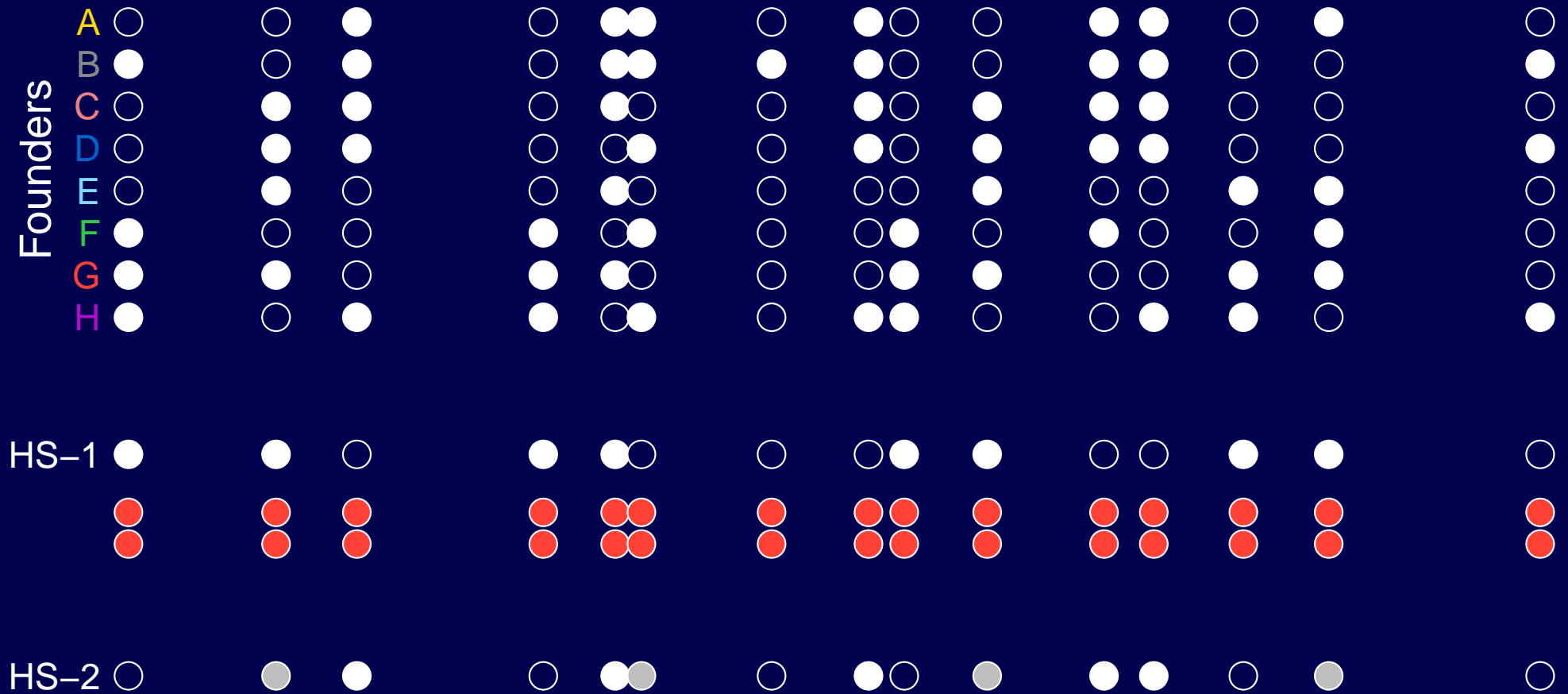1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X

HS genotype reconstruction

20

HS genotype reconstruction